MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Univariate and Multivariate Long Short Term Memory (LSTM) Model to Predict Covid-19 Cases in Malaysia Using Integrated Meteorological Data

Ng Wei Shen, Azuraliza Abu Bakar*, Hazura Mohamad

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract The rate of transmission of coronavirus disease (COVID-19) has been very fast since the first reported case in December 2019 in Wuhan, China. The disease has infected more than 3 million people worldwide and resulted in more than 224 thousand deaths as of May 1, 2020, reported by *The World Health Organization* (WHO). In the past, meteorological parameters such as temperature and humidity were essential and effective factors against serious infectious diseases such as influenza and Severe Acute Respiratory Syndrome (SARS). Therefore, exploring the relationship between meteorological factors and active COVID-19 cases is essential. This study employs the *LONG-SHORT TERM MEMORY* (LSTM) method to predict Covid-19 Cases in Malaysia. We propose a univariate and multivariate model using Covid-19 cases and meteorology data. The univariate LSTM model uses Covid-19 active cases data in a year as a control attribute for model development. The multivariate LSTM model uses the integrated Covid-19 cases, and meteorology data consists of attributes: minimum, maximum, and average values of Humidity, Temperature, Windspeed, and Pressure from 13 states of Malaysia. The model's performance is evaluated using errors such as MAE, RMSE, MAPE, and the R2 Score. The low errors and higher R2 score indicate the model's excellent performance. We observed that the univariate LSTM model gives the least error in five states, indicating that those states' daily active cases are the main contributing factors. In the multivariate LSTM model, the daily cases and humidity, temperature, and windspeed are the main factors in several different states. The result of the study is to help the government to prevent and manage the spread of the COVID-19 and other upcoming pandemic better.

Keywords: Long Short Term Memory, Univariate and multivariate model, Active covid-19 cases, meteorology.

Introduction

In early December, the severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) variant called COVID-19 appeared in Wuhan, China. Since the coronavirus was discovered, the virus has quickly spread worldwide and makes local transmission in many countries, including America, Europe, Africa, and Asia. Understanding the transmission patterns of Covid-19 could slow down the rapid spread of the disease. On January 30, 2020, the *World Health Organization* (WHO) (2020) [18] designated the COVID-19 pandemic as a public health emergency that required attention and confirmed the pandemic as a global pandemic on March 11, 2020. Like other countries, COVID-19 cases have also been reported in Malaysia. On February 4, 2020, the first COVID-19 case was detected; A senior citizen recently returned from Singapore with a fever and cough. In Malaysia, 26 thousand COVID-19 cases and 1130 deaths have been documented as of February 2021.

Environmental factors influence the epidemiology of most infectious diseases. Several studies have found that climatic and weather conditions can affect the distribution of regional and temporal infectious diseases (Valsamatzi-Panagiotou & Penchovsky, 2021). The coronavirus family, such as SARS CoV-1

***For correspondence:**
azuraliza@ukm.edu.my

Received: 23 Oct. 2022
Accepted: 27 July 2023

10.11113/mjfas.v19n4.2814

and MERs CoV, have found that the viruses vary seasonally and prefer low temperatures and humidity (Pal, *et al.,* 2020). Additionally, in the early phase of the COVID-19 pandemic, researchers reported that temperatures had a positive correlation and humidity had a negative correlation with cases worldwide (Bashir, *et al.,* 2020). However, a negative linear association exists between temperature and cumulative COVID-19 cases (Prata *et al*., 2020)]. Numerous studies have suggested that the spread of COVID-19 occurs more in cold and temperate climates than in hot and tropical climates, in line with the behavior of seasonal flu respiratory viruses by Bloom-Feshbach *et al.,* (2013). Meteorological data such as temperature and humidity were also used as several earlier studies have shown that they are correlated with the spread of the pandemic.

Deep learning specifically the Long Short Term Memory (LSTM) algorithm has been successfully used to anticipate dengue and *influenza* outbreaks by Leonenko *et al.,* (2017) [7]. Furthermore, previous research has examined whether relative humidity and absolute humidity are critical in transmitting COVID-19. At the same time, studies on *influenza* viruses reveal that specific slowdowns are important aspects of viral transmission. The COVID-19 virus is believed to be transmitted directly through respiratory droplets from other people who are contagious to healthy people within a one-meter distance. In addition, healthy people can also be infected indirectly through contaminated surfaces by *World Health Organization* (WHO) (2020) [19]. However, research on the factors affecting the spread of the COVID-19 virus is sorely lacking, which may be one factor that makes the COVID-19 virus irresistible. The study on the impact of weather variables and COVID-19 transmission is critical.

This study employs a deep learning algorithm using integrated meteorology and the Covid-19 cases data in Malaysia. The long short-term memory (LSTM) algorithm is developed to learn the univariate and multivariate time series weather and Covid-19 case data for Malaysia's states. We evaluated the impact of weather on active COVID-19 case patterns and examined the specific humidity and other climate parameters affecting COVID-19 transmission and forecasting in Malaysia.

Related Work

Insufficient information on COVID-19 has made it more difficult for the world to handle its continuous implosion. In both laboratory and epidemiological studies, meteorological and environmental factors have been reported to affect the survival and transmission of the virus. Active research has been conducted to find the effect of meteorological factors on the trend of the virus.

Several studies explore the Covid-19 case data concerning Malaysia's meteorological and air quality data. The meteorological parameters are temperature, humidity, wind speed, humidity, and air quality data such as PM2.5, PM10, NO2, and O3. Jalaludin *et al*. (2023), Mohan *et al*. (2022), Suhaimi, *et al.* (2020) Jalaludin *et al*. (2023) highlight the significant correlation between Malaysia's air pollution, meteorological parameters, and COVID-19 cases. The findings indicate that COVID-19 cases were positively correlated with O3, NO2, RH, PM10, and PM2.5 but negatively correlated with Solar radiation (SR) and Windspeed (WS).

Mohan *et al*. (2022) reveal that the influence of meteorological factors such as temperature (T) between 23 and 25 °C and relative humidity (RH) (70–80%) in the pandemic spread and air quality to the spread of Covid 19 cases by an increase in the infected cases in northern and central Peninsular Malaysia. They gathered COVID-19 data on meteorological parameters and air quality index (AQI) during three movement control order (MCO) periods covering the state of Selangor, Kuala Lumpur and Putrajaya. Suhaimi *et al*. (2020) determined associations between air quality, meteorological factors, and COVID-19 cases in Kuala Lumpur, Malaysia. Air pollutants and meteorological data in 2018–2020 were obtained from the Department of Environment Malaysia, while daily new COVID-19 cases in 2020 were obtained from the Ministry of Health Malaysia. The study demonstrates the remarkable connection between air pollution, meteorological factors, and susceptibility to COVID-19 infections in Malaysia. Makama & Lim (2021) employed a generalized additive model (GAM) on ambient temperature and absolute humidity with the new daily COVID-19 infection (NDI). They found a positive association between *temperature and covid cases*, particularly above 29.7°C. In contrast, the association with absolute humidity showed a stronger positive relationship below 22.6 g/m3, indicating that COVID-19 could not be suppressed in warmer weather.

The discussion above covers studies on the effects of the meteorological and air quality data on Covid-19 cases in Malaysia. The studies mainly cover congested areas such as Selangor, Kuala Lumpur, and Putrajaya. All studies revealed the influence of meteorological data on Covid cases. This paper enhances the above study by employing a deep learning algorithm and covers all states in Malaysia.

The use of the time series analysis method, the Auto-Regressive Integrated Moving Average (ARIMA), to determine the factors contributing to the Covid-19 cases has shown some promising results. The ARIMA model parameters are the autoregression sequence, the degree of flow difference, and the Moving Average order. The correlation analysis shows a negative correlation between the number of cases and temperature, which means the number of cases will rise if the temperature in a particular area is low. Additionally, the study by Rendana & Idris (2021) concluded that sunlight, rain, and temperature factors could significantly contribute to COVID-19 cases.

The short-long term memory algorithm (LSTM) is one of the Recurrent Neural Network (RNN) types that is actively investigated for time series data analysis. This method is proposed because LSTM is ideal for classifying, processing, and predicting time series given the unknown time intervals of its duration. LSTM has an advantage over the RNN model, *the Hidden Markov* model, and other sequential learning approaches due to its insensitivity to the length of the gap by Miguel-Hurtado *et al.,* (2016) [9]. The LSTM can learn the daily active COVID-19 cases and meteorological data to find the correlation between the data and can be used to predict active COVID-19 cases later.

Bakar *et al*. (2022) developed the air quality model based on the LSTM and ARIMA to predict the particulate matter 10 micrometers or less in diameter (PM10) in Malaysia. They used the air quality data from the Department of Environment Malaysia from July 2017 to June 2019. The results showed that forecasting for PM10 using the multivariate LSTM model gave lower RMSE than the univariate LSTM model for those selected stations. Shrivastav & Jha (2021) used the Gradient Boosting Model (GBM) to examine the impact of temperature and humidity on COVID-19 virus transmission rates in India. GBM is an improved tree version that removes the weak predictors and selects a stronger one. The model has efficient distributed characteristics and an environment for model tuning and selection. The GBM model is optimized for 50 randomly selected trees and tuned with the number of trees, learning rate, number of folds, and distribution function. GBM suggests better Poisson distribution performance in the second forecast of two active and recovered COVID-19 cases.

The study by Bhimala *et al.* (2021) showed the capability of the univariate and multivariate LSTM in predicting the COVID-19 cases related to weather data in 28 states in India. The study used the confirmed COVID-19 cases from April 1 to July 31, 2020. The daily meteorological parameters of the specified period consist of temperature (minimum, maximum and mean) and specific humidity (SH) extracted from NCEP. The correlation coefficient between specific humidity and COVID-19 cases is critical in the area. This study shows that LSTM model forecasting is enhanced on medium and long-distance scales due to weather data integration in India.

According to Rashed& Hirata (2021), public mobility is a dominant factor in estimating new positive cases, and meteorological data improve prediction accuracy. They developed the multi-path LSTM model to predict the spread of COVID-19. The model was tested using different time frames, and the results were compared to Google Cloud forecasts. The model gives significant improvement compared with Google Cloud forecasting. This model can provide public awareness regarding the morbidity risk of the COVID-19 pandemic in a feasible manner.

Past studies have shown the capability of deep learning in predicting the COVID-19 cases dataset integrated with the weather data. This paper presents the univariate and multivariate LSTM model on integrated COVID-19 for Malaysian Covid-19 cases. The use of machine learning in virus diagnosis is rising in health. Long Short-Term Memory (LSTM) is one of the machine learning algorithms of the recurrent neural network (RNN) that attempts to model time or sequence of data. LSTM is a deep learning algorithm giving the network layer at time t to input from the same network layer at time t+1. The proposed LSTM was implemented using on Keras platform (Figure 1).

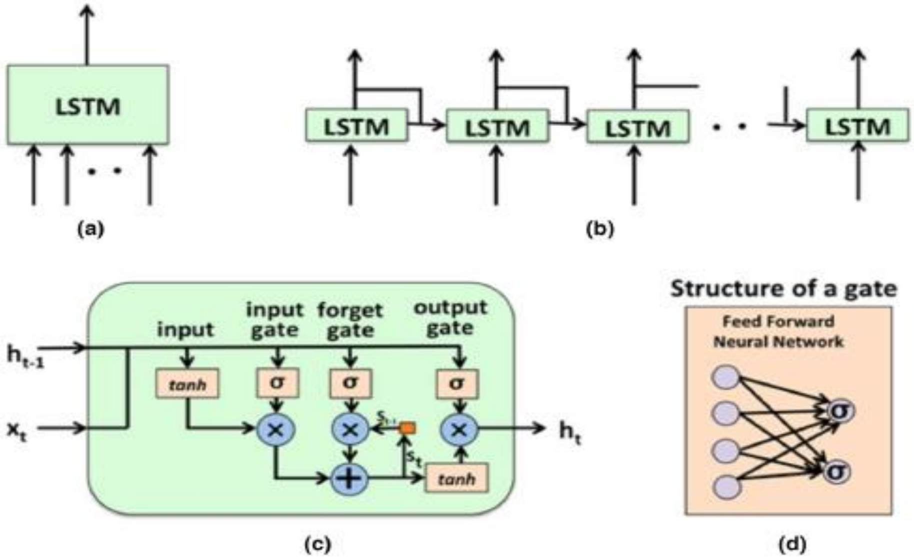


Figure 1. Keras Implimentation of LSTM with multiple parameters (a) Basic LSTM Structure (b) unfolded LSTM Representation (c) LSTM cell architecture (d) Internal structure of cell doors (M. Rastogi, 2020).

The LSTM network block diagram of multiple basic inputs and memory transformations between each LSTM cell is presented in Figures 1a and b. LSTM cells are made up of three doors: input door (*input gate, it)*, forget door (*forget gate, ft*) and output door (*output gate, ot*) with different functions shown in Figure 1c. The forgets gate to be responsible for forgetting unnecessary information, while input gate is used to add new useful information. The output gate updates the hidden state at every step of the time. Each gate is a front-pointing network with many hidden units as shown in Figure 1d. The mathematical representation of LSTM is given under Equation 1 – 4 where the σ, i, f, o, and g represent *the sigmoid* function, the input door, the forget gate, the output door and the transformation of the ungated input respectively. Weights ($w_i, w_f, w_o, w_g, and\ u_i, u_f, u_o, u_g$) represented in matrix format, bias (b_i, b_f, b_o, b_g) are represented by vectors and s_{t-1} represent the cell state in the previous (M. Rastogi, 2020).

$$i_t = \sigma(w_i x_t + u_i h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma\left(w_f x_t + u_f h_{t-1} + b_f\right) \quad (2)$$
$$o_t = \sigma(w_o x_t + u_o h_{t-1} + b_0) \quad (3)$$
$$h_t = o_t \times tanh\ tanh\ \left(i_t \times tanh\ tanh\ \left(w_g x_t + u_g h_{t-1} + b_g\right) + f_t \times s_{t-1}\right) \quad (4)$$

Past studies have shown the capability of deep learning in predicting the COVID-19 cases dataset integrated with the weather data. The project aims to use the *Long Short-Term Memory* (LSTM) algorithm to produce integrated COVID-19 predictions of meteorological data and develop machine learning models that can produce accurate results.

Materials and Methods

This study uses four phases methodology; data understanding, data preparation, model development, model evaluation. These phases are part of the CRISP-DM data mining methodology [6].

Dataset

The data source used for the study was Malaysian COVID-19 cases data from the Ministry of Health Malaysia's GitHub repository website (https://github.com/MoH-Malaysia/covid19-public), a total of 5161 records of daily active COVID-19 case data in the period October 2020 to October 2021 were extracted for this study. Meteorological data consists of twelve parameters; average, maximum, and minimum temperature, humidity, wind speed, and pressure. The data were obtained from https://www.wunderground.com and https://www.timeanddate.com/. It consists of the time and date of daily active Covid-19 cases in thirteen Malaysian states from October 2020 to October 2021. Table 1 shows the average active cases in states for the duration of one year.

Table 1. Average active cases (by state) October 2020 - October 2021

State	Average Active Cases	State	Average Active Cases
Melaka	2043.73	Perlis	153.53
Pahang	2489.06	Penang	3862.52
N.Sembilan	2827.52	Sabah	6078.63
Kelantan	3494.74	Terengganu	1728.32
Sarawak	7046.69	Perak	2862.55
Johor	7300.95	Kedah	3543.87
		Selangor	22289.02

Data Preparation

The data preparation phase involves data integration and cleaning. Due to daily COVID-19 positive case data and meteorological data being from different sources, both data sources were integrated according to date and state. Data exploration was carried out to ensure an understanding of the form of the data set before being used to develop the forecast model. There were 5161 lines of daily active COVID-19 case data, and meteorology was collected in this study with 15 characteristics. The process of selecting essential parameters is carried out to select important attributes for developing an accurate prediction model. We employed the decision tree classifier specifically the Gini Index to select the important features in the dataset. Figure 2 demonstrates the chart of the attribute importance.

The attribute importance presented in Figure 2 shows that the mean humidity (humidity_mean) has the highest importance in Malaysia's daily active COVID-19 cases. Meanwhile, the attributes mean wind speed (windspeed_mean) and mean temperature (temperature_mean) have relatively similar significance for that data set, followed by the maximum wind speed (windspeed_max), which reaches a value of 0.12176. Attributes maximum temperature (temperature_max) and minimum temperature (temperature_min) have reached the values of 0.09518 and the value 0.07714. In contrast, the maximum humidity(humidity_max) attribute has reached the value of 0.06416, indicating that these attributes are attributes of moderate importance to the data set. For attributes maximum, mean, and minimum pressure (pressure_max, pressure_mean, pressure_min, and windspeed_min) have a value of interest less than 0.05 and indicate that these attributes are not crucial for developing such predictive models. Therefore, these attributes will be removed from developing the daily COVID-19 active case prediction model.

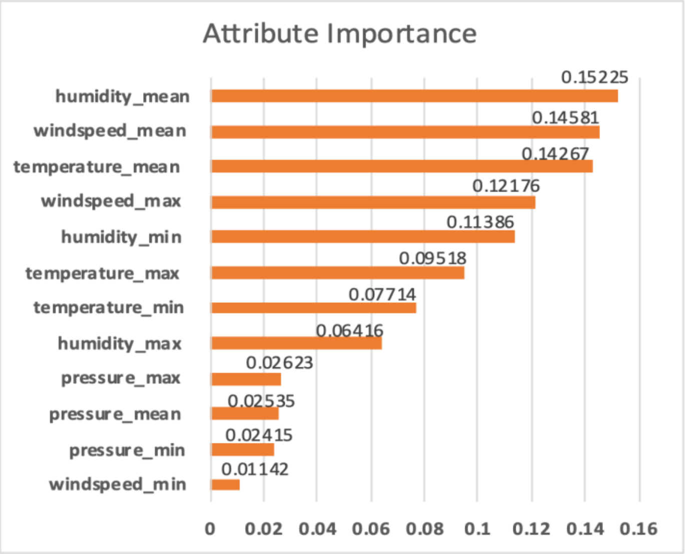


Figure 2. Attribute Importance Ranking

Model Development

The algorithm used in this project is the Long Short-Term Memory (LSTM) algorithm. We refer the work by Bhimala *et al.* (2021) as the baseline. The LSTM algorithm is performed by using the output of the neural network layer at the time t as the input for the same network layer at the time t+1. The LSTM algorithm is a well-known algorithm often used on predictive models. The LSTM algorithm has been successfully used to anticipate dengue and influenza outbreaks. We employed two LSTM models, the univariate and multivariate LSTM, to predict daily COVID-19 cases for each state in Malaysia. The time series data from 1 October 2020 to 31 October 2021 was selected for the study and divided into two parts; the first eleven months' data were selected as training sets, while the remaining month's data were made as test sets. The LSTM univariate model used the Covid-19 daily cases data as the control trials set (CTLs). The multivariate LSTM modeling will use the combination of CTL control data and individual meteorology attributes (humidity, temperature, wind speed, pressure) to investigate the impact of weather on the transmission of the COVID-19 virus and cases. The univariate and multivariate LSTM models have been optimized with minimal error conditions for prediction. Next, prediction models that combine different weather parameters are generated and evaluated with data observed for high prevalence conditions for COVID-19 cases in Malaysia.

Model Evaluation

This study evaluates the LSTM model to determine the relationship between meteorology attributes and the Covid-19 daily cases. The evaluation metrics include the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 score to evaluate the model's performance in regression analysis as in Kelleher, *et al.,* (2015). The MAE represents the average of the absolute difference between the actual and predicted values in the dataset to measure forecast accuracy. It measures the dataset's average residuals (See Eq. 5).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{y} \qquad (5)$$

where $\hat{y}$ is the predicted value for the mean of y_i and N is the length of the observed data.

RMSE is the square root of the mean squared error. It measures the standard deviation of residuals. See Eq. 6

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad (6)$$

The MAPE evaluation metric is an error metric used to measure the performance of a regression machine learning model. MAPE is one of the most commonly used KPIs to measure forecast accuracy. MAPE is the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors. The lower value of the MAPE means that the model's performance is better. See Eq. 7

$$MAPE = \frac{1}{N}\sum_{t-1}^{N}\left|\frac{(y_i - \hat{y})}{d_t}\right| \tag{7}$$

R^2 is the coefficient of determination representing the proportion of the variance in the dependent variable, which the linear regression model explains. It is a scale-free score; irrespective of the values being small or large, the value of R^2 will be less than one, which is in the range of [0,1]. The closer the fitting coefficient to an excellent value of 1, the more accurate the prediction model. R^2 are used to explain how well the independent variables are in the linear regression model and the variability in the dependent variable. As for the R^2 value, a high value indicates that the variance of the model is similar to the actual value, while a low R^2 value indicates that the two values are not closely related See Eq. 8.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2)} \tag{8}$$

The lower value of MAE, MAPE, and RMSE implies higher accuracy of a regression model, indicating high forecast precision by Alsayed (2020). However, a higher value of R^2 is considered desirable.

Results and Discussion

LSTM univariate and multivariate models were developed to predict active COVID-19 cases for each state in Malaysia. The LSTM multivariate model was developed to improve prediction accuracy in the daily COVID-19 active case in respective states in Malaysia to find the most contributing meteorology factors that affect the daily COvid-19 cases. The study used the data set for 1 October 2020 to September 2021 for training and the dataset of October 2021 as a test set, while the other data set was used as a training set.

Univariate LSTM Prediction Model

Figure 3 showed the prediction performance of LSTM univariate models for 13 states in Malaysia. The prediction relative values (CTL) to the actual value of October 2021 test data. The red line represents the prediction values while the grey line represents the actual values.

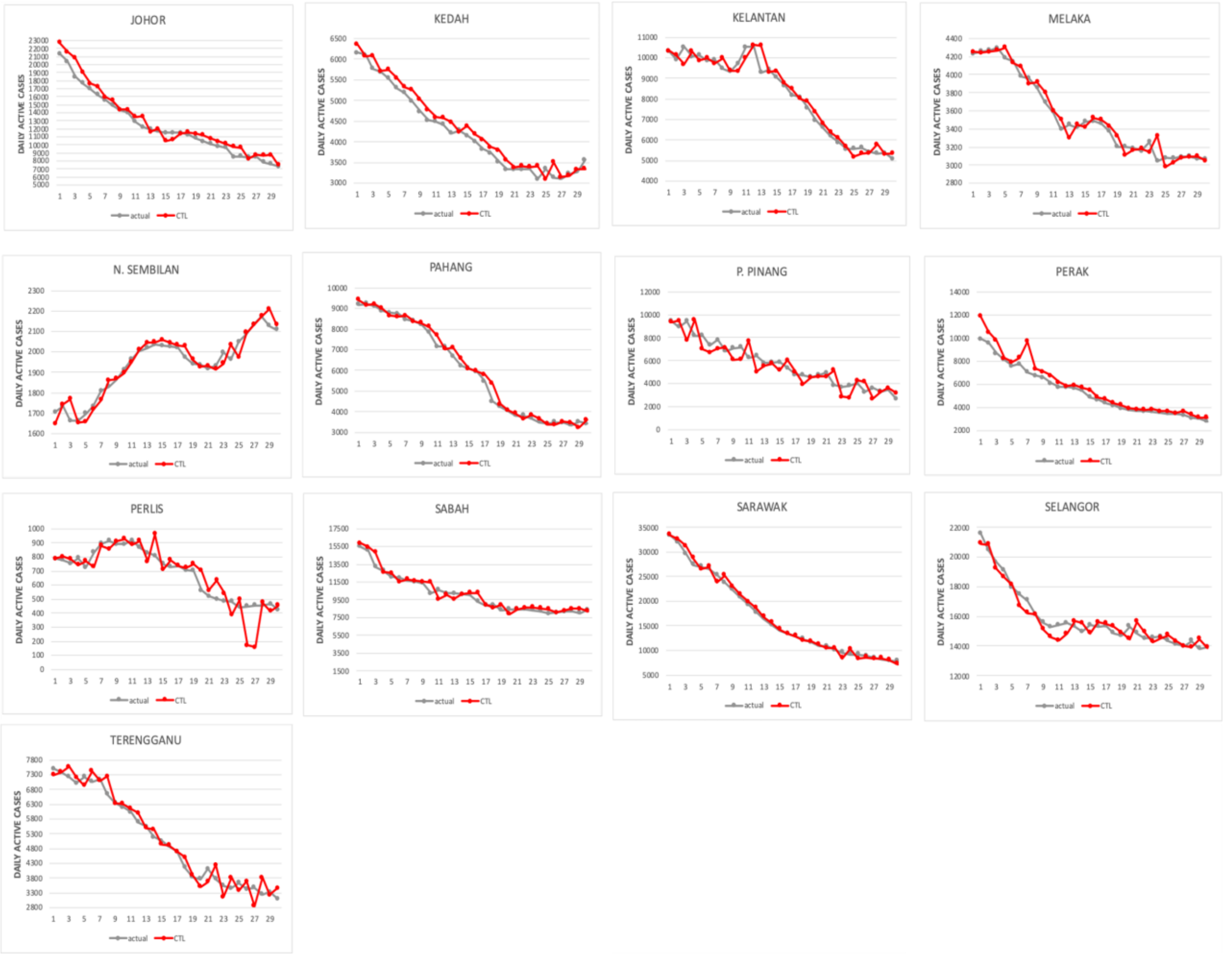


Figure 3. Prediction Covid-19 Daily Cases for Univariate LSTM Model (Predicted vs Actual Values)

The univariate LSTM model for each state and the values for MAE, MAPE, RMSE and R^2 of the training data are shown in Table 2. The performance results of the univariate LSTM model were used to predict daily active cases of COVID-19 in Malaysia. Figure 4 shows the performance results in the form of a bar graph (by state).

Table 2. Performance results of the univariate LSTM model

	MAE	RMSE	MAPE	R^2
N. Sembilan	29.44	39.695	1.56	0.932
Melaka	63.864	85.373	1.86	0.962
Perlis	95.996	95.996	11.66	0.674
Kedah	173.213	201.056	4.25	0.961
Pahang	179.645	247.705	3.38	0.988
Kelantan	291.999	388.478	3.62	0.961
Terengganu	300.994	300.994	5.58	0.962
Johor	373.346	484.344	3.28	0.986
Selangor	436.568	512.711	2.75	0.947
Sabah	531.455	531.455	3.77	0.946
Sarawak	719.303	719.303	3.69	0.993
Perak	749.734	749.734	8.08	0.883
P.Pinang	860.186	860.186	12.77	0.827

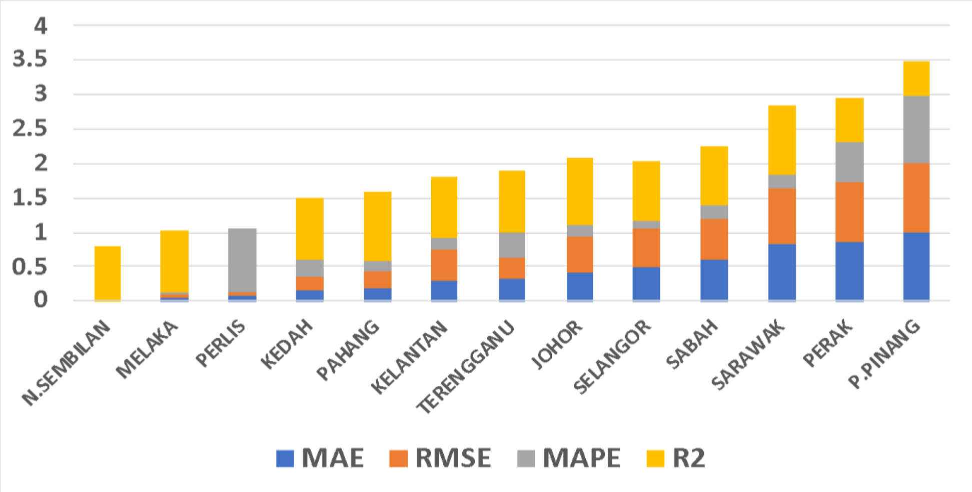


Figure 4. Performance Results of the Univariate LSTM (by state)

Based on Table 2 and Figure 4, the MAE value of Negeri Sembilan is the lowest, followed by the state of Melaka and Perlis. The MAE value comparison graph showed that the performance of the LSTM Univariate model for predicting daily active cases of COVID-19 in Negeri Sembilan was the highest by achieving an MAE value of 29.44, followed by the state of Melaka, which reached 63.864 and the state of Perlis which reached 95.996. MAE for Selangor, Sabah, Sarawak, Perak, and P. Pinang are 436.568, 531.455, 719.303, 749.734, and 860.186, respectively. The values are more than 50% in the range of 29.44 and 860.186. These values show that univariate LSTM is not a good model to predict the performance of those states, especially P. Pinang. A lower MAE value means that good model performance.

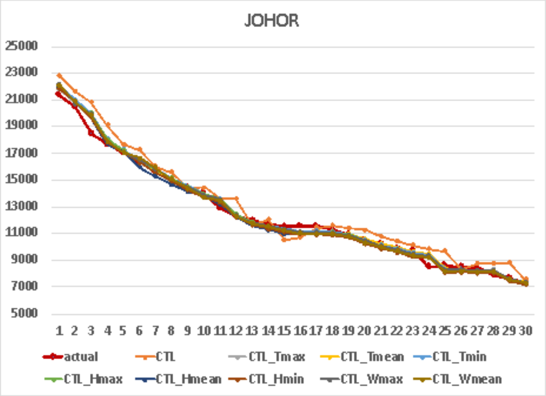
The RMSE value comparison graph showed that the performance of the LSTM Univariate model for Negeri Sembilan was the lowest at 39.695, followed by Melaka, which achieved 85.373, and Perlis, with 95.996. The result showed that the model performance of the three states was the same as the previous MAE results. RMSE for Johor, Selangor, Sabah, Sarawak, Perak, and P. Pinang are 484.344, 512.711, 531.455, 719.303, 749.734 and 860.186, respectively. These values show that univariate LSTM is low in the accuracy of those states, especially P. Pinang. The values are more than 50% in the range of 39.695 and 860.186. A lower RMSE value indicates a higher accuracy in predicting performance.

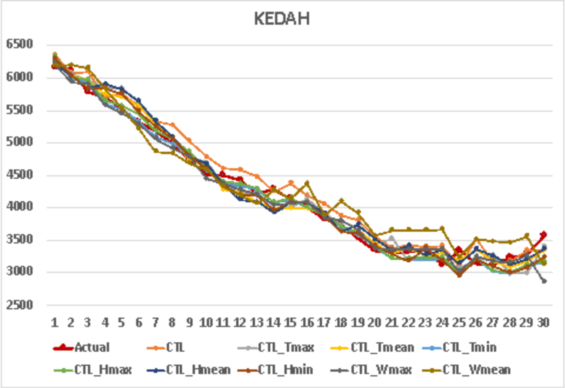
The MAPE value comparison graph shows that the model performance for Negeri Sembilan was the lowest at 1.56%, followed by Melaka, which achieved 1.86%. Only both states achieved a low MAPE value of 2%, and the state of P. Pinang achieved a MAPE value of 12.77%. Based on the MAE, RMSE, and MAPE assessment metrics, it can be found that the LSTM Univariate model is not suitable for predicting active COVID-19 cases in the state of Penang.

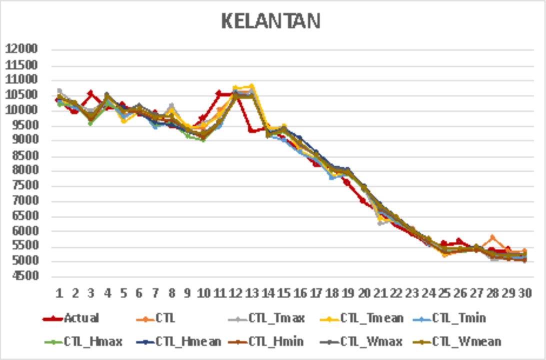
The graph shows that the R^2 value for Sarawak was the highest at 0.993; Pahang and Johor were at 0.988 and 0.986, respectively. The state of Perlis achieved the lowest R^2 at 0.674, although small error values in MAE and RMSE. Overall, it can be seen that N. Sembilan and Melaka give the best prediction capability with the lowest error values (MAE, RMSE, and MAPE) and high R^2 values, which indicates a better model obtained.

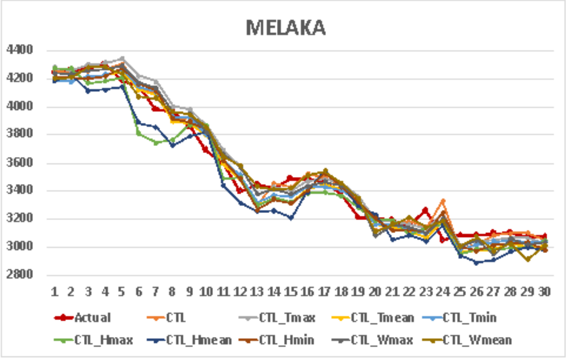
Multivariate LSTM Model

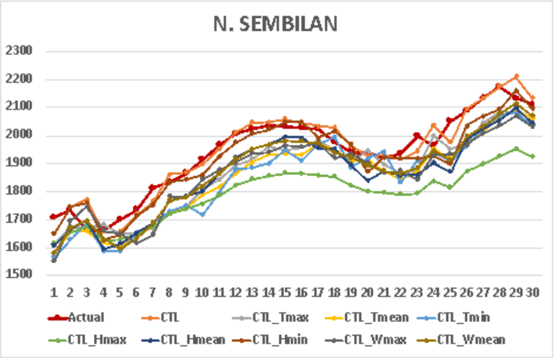
LSTM Multivariate model was developed to improve the prediction performance in the daily COVID-19 active case forecast for states in Malaysia. Figure 5 presents each state's forecast results of the multivariate LSTM model during the test period (1 October to 30 October 2020) for the 13 states in Malaysia, where L1 to L30 represents the 1 to 30 days of lag data utilized for forecasting the next day COVID-19 cases. In the experiment of the multivariate model, we represent the Covid-19 cases vs original meteorology attributes as follows: humidity_mean (CTL-Hmean), windspeed_mean (CTL_Wmean), temperature_mean (CTL_Tmean), windspeed_max (CTL_Wmax), humidity_min (CTL_Hmin), temperature_max (CTL_Tmax), temperature_min (CTL_Tmin), humidity_max (CTL_Hmax), windspeed_min (CTL_Wmin).

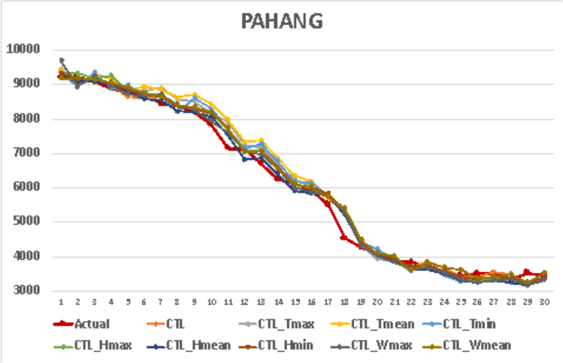

JOHOR
25000
23000
21000
19000
17000
15000
13000
11000
9000
7000
5000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean


KEDAH
6500
6000
5500
5000
4500
4000
3500
3000
2500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean


KELANTAN
12000
11500
11000
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean


MELAKA
4400
4200
4000
3800
3600
3400
3200
3000
2800
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean


N. SEMBILAN
2300
2200
2100
2000
1900
1800
1700
1600
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean


PAHANG
10000
9000
8000
7000
6000
5000
4000
3000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Actual
CTL
CTL_Tmax
CTL_Tmean
CTL_Tmin
CTL_Hmax
CTL_Hmean
CTL_Hmin
CTL_Wmax
CTL_Wmean

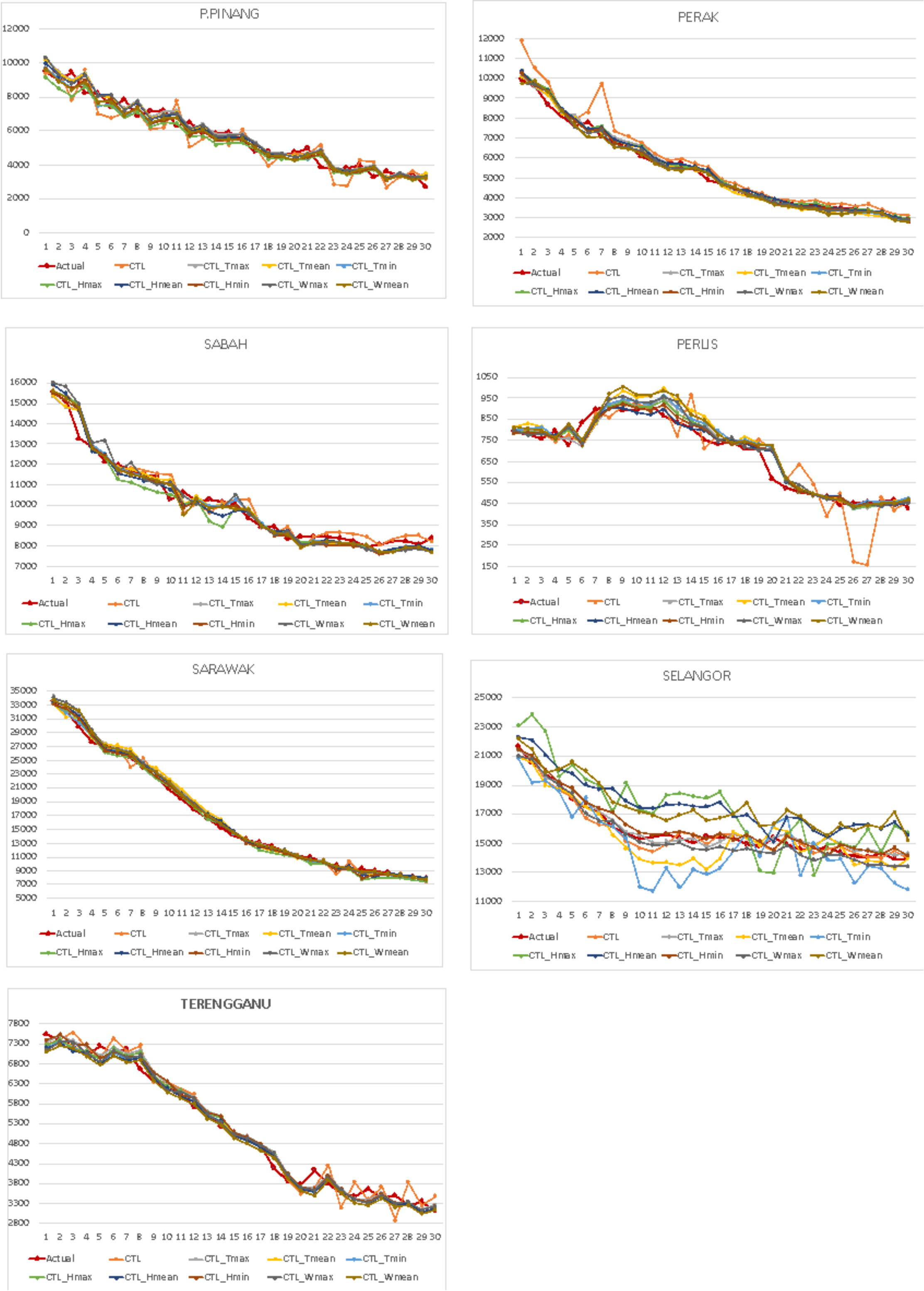


Figure 5. Prediction of Covid-19 Cases of Univariate (CTL) and Multivariate LSTM Models (Minimum, maximum and mean temperature (T), humidity(H), and windspeed(W))

To find the most contributing meteorological factors in Covid-19 cases, we rank the values of MAE, RMSE, MAPE, and R^2 and observe the top four ranks (Least MAE, RMSE, and MAPE values but largest R^2 value) as shown in Figure 6. The figure shows that the univariate LSTM model outperformed the multivariate LSTM model for Johor, Kelantan, Melaka, Negeri Sembilan, Pahang, and Sarawak. Thus, there is a weak correlation between meteorological attributes and daily active cases of COVID-19 in these states. The Multivariate LSTM model achieved higher performance than the Univariate LSTM model in Kedah, Perak, P. Pinang, Perlis, Sabah, Selangor, and Terengganu. The Multivariate LSTM model indicates a stronger correlation between meteorological attributes and daily active cases in several states. Temperature affects the cases in Kedah, Perak, Selangor, and Terengganu, while humidity affects Penang, Perlis, and Sabah. There is no correlation between the states' average active cases with the factors affected (refer to Table 1). However, it can be seen that temperature affects the larger states, while humidity affects small states in the northern part of Malaysia (Perlis, Penang, Sabah). Generally, the meteorology factors did not affect the southern part of Malaysia in the first rank.

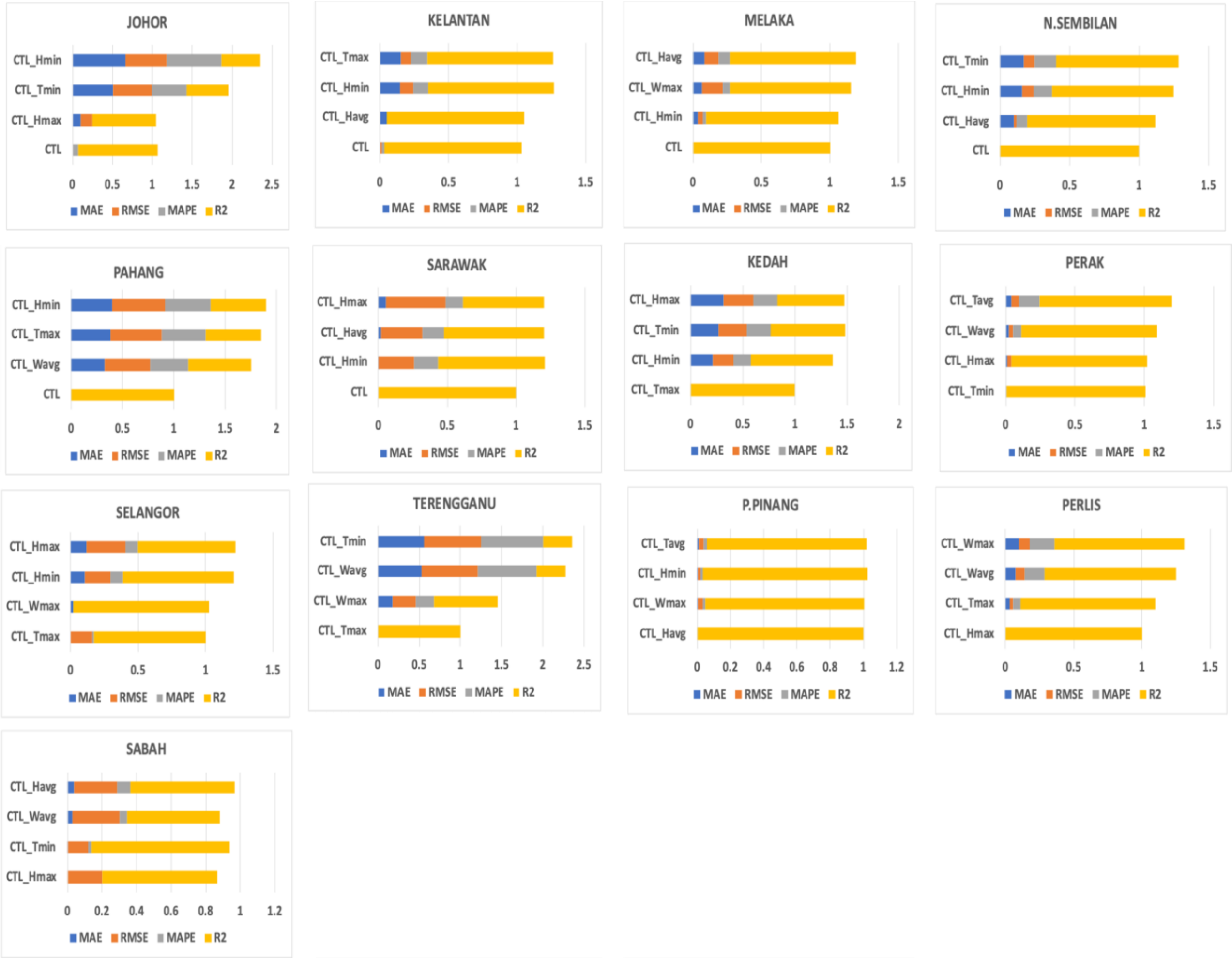
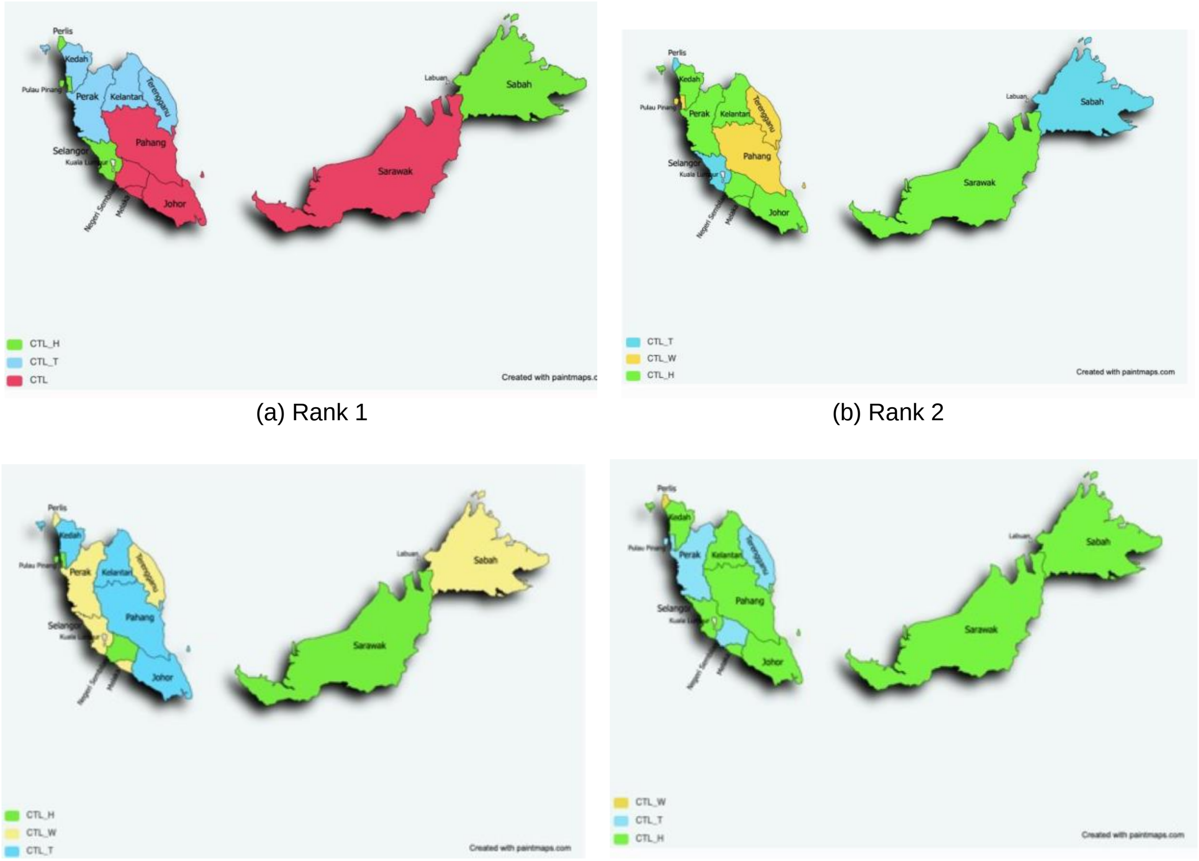


Figure 6. Attribute ranking by states for multivariate LSTM

We illustrate the top four ranks of the attributes for each state in Figure 6 in Malaysia map as in Figure 7. The maximum, minimum, and average cases of humidity(H), temperature(T), and Windspeed(W) are generalized to CTL-H, CTL-T, CTL-W respectively, and the specific daily cases as CTL. The study observed that in Rank 1, the specific daily cases contribute to Malaysia's south and eastern states. The temperature affects the four states in the northern part of Malaysia, while the humidity covers the rest of the four states. In Rank 2, it can be observed that the humidity factors dominated seven states, and the temperature and windspeed affected the other three states, respectively. In Rank 2 and 3, it can be seen that humidity affects the Covid-19 cases in Sarawak and Negeri Sembilan. In Rank 4, the humidity dominates Sabah and Sarawak and most states in Peninsular Malaysia.



(a) Rank 1

(b) Rank 2

(c) Rank 3

(d) Rank 4

Figure 7. The attribute importance ranked in four levels (a-d) illustrated in Malaysia map (CTL: Covid-19 Daily Cases; CTL-H: Cases vs Humidity; CTL-T: Cases vs Temperature; CTL-W: Cases vs windspeed)

We describe our finding in general compared to three similar studies by Rashid & Hirata (2021), Shrivastav & Jha (2021) and Bhimala *et al*. (2021) employing deep learning in meteorological and Covid-19 case data. A numerical comparison could not be provided since the study involved different countries, geographical locations, and climates. The findings in this paper are consistent with their finding that revealed the specific humidity and temperature are among the main factors influencing the cases. In all studies, the forecasting patterns were almost consistent with the actual data spreading throughout the timestamps. Different factors at different ranks influenced Covid cases in different states of Malaysia due to different climatic characteristics.

Conclusion

This study investigates the effects of meteorological data on Covid-19 cases in Malaysia. We employed the univariate and multivariate LSTM algorithms to forecast the Covid-19 cases in one year. The findings showed that the univariate model, which uses only the covid-19 cases, affects certain states in Malaysia, mainly in the southern part of the country. The temperature factors affect most in the larger states, and the humidity affects the smaller states in northern Malaysia. Considering the ranks of the factors, we revealed that, in the first rank, different states give different contributing parameters to the covid-19 cases. However, the overall results showed that humidity and temperature dominate all states after the fourth rank. The windspeed has a slight effect in several states, such as Pahang and Terengganu. Our results suggested that the univariate LSTM model, which includes the confirmed COVID-19 time series data, outperformed the southern and middle parts of Malaysia such as Johor, Kelantan, Melaka, Negeri Sembilan, Pahang, and Sarawak as the top rank of the feature. The result also showed that in multivariate LSTM obtained humidity, temperature, and windspeed as the factors that affect the Covid-19 cases within training in the period of 1 October 2020 to September 2021 and the test period of 1

October to 30 October 2021 in the top four ranks of features. This study leads to alerting the government for ruling a specific initiative and preparation for pandemic management.

Deep learning technology in health care helps doctors analyse diseases and better treat a particular disease. So, the medical decisions made by doctors can be made more wisely and are improving in standards. The deep learning model's predictions give critical insights into the input data and learned characteristics, enabling knowledge analysis and decision-making. An efficient predictive model could lead to effective government and public health interventions, and adequate healthcare capacity has helped reduce COVID-19 cases in Malaysia. The meteorology factors identified in this paper could also alert the responsible department to alert the country for the possibility of the pandemic hitting certain seasons. Our future research will include integrating the additional meteorological with the air quality parameters to identify the influences of both parameters in the transmission of COVID-19 on atmospheric conditions by using an efficient and robust deep-learning approach to improve the forecasting performance of deep-learning methods.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

This work is part of a research project, supported by the Universiti Kebangsaan Malaysia.

References

[1] Alsayed, A., Sadir, H., Kamil, R., Sari, H. (2020). Prediction of epidemic peak and infected cases for COVID-19 disease in Malaysia. *Int J Environ Res Public Health. 17*(11):1–15. https://www.mdpi.com/1660-4601/17/11/4076.

[2] Bakar, A., Aftar, M. Ariff, M. Nadzir, M. A. Shahrul, M., Aftar, M., Wen, Ong, Suris, Fatin Nur Afiqah. (2022). Prediction of multivariate air quality time series data using long short-term memory network. *Malaysian Journal of Fundamental and Applied Sciences. 18*, 52-59. 10.11113/mjfas.v18n1.2393.

[3] Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., Bashir, M. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA*. The Science of the Total Environment*, *728*, 138835. https://doi.org/10.1016/j.scitotenv.2020.138835.

[4] Bhimala, K. R., Patra, G. K., Mopuri, R., Mutheneni, S. R. Prediction of COVID-19 cases using the weather integrated deep learning approach for India. (2021). *Transboundary and Emerging Diseases. 69*(32022Pagesi- 927-1662. Doi: 10.1111/tbed.14102.

[5] Bloom-Feshbach, K., Alonso, W. J., Charu, V., Tamerius, J., Simonsen, L., Miller, M. A., Viboud, C. (2013). Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): A global comparative review. *PLoS ONE, 8*(2), e54445. https://doi.org/10.1371/journal.pone.0054445.

[6] Heidari, A., Jafari Navimipour, N., Unal, M. *et al*. (2022). Machine learning applications for COVID-19 outbreak management. *Neural Comput & Applic., 34*, 15313-15348. https://doi.org/10.1007/s00521-022-07424-w.

[7] Jalaludin, J., Wan Mansor, W. N., Abidin, N. A., Suhaimi, N. F., Chao, H-R. (2023). The impact of air quality and meteorology on COVID-19 cases at Kuala Lumpur and Selangor, Malaysia and prediction using machine learning. *Atmosphere*, *14*(6), 973. https://doi.org/10.3390/atmos14060973.

[8] Kelleher, J. d., Namee, M. B. & D'Arcy, A. (2015). *Fundamental of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies*. The MIT Press.

[9] Leonenko, V. N., Bochenina, K. O., & Kesarev, S. A. (2017). Influenza peaks forecasting in Russia: Assessing the applicability of statistical methods. *Procedia Computer Scienc*e, *108*, 2363-2367. 10.1016/j.procs.2017.05.196.

[10] Makama, E. K. and Lim, H. S. (2021). Effects of location-specific meteorological factors on COVID-19 daily infection in a tropical climate: A case of Kuala Lumpur, Malaysia*. Hindawi Advances in Meteorology*. https://doi.org/10.1155/2021/6675943.

[11] Manu Rastogi. (2020). Tutorial on LSTMs: A Computational Perspective. https://towardsdatascience.com/tutorial-on-lstm-a-computational-perspective-f3417442c2cd#0d00.

[12] Miguel-Hurtado, Oscar & Guest, Richard & Stevenage, Sarah & Neil, Greg & Black, Sue. (2016). Comparing machine learning classifiers and linear/logistic regression to explore the relationship between hand dimensions and demographic characteristics. *PLoS ONE*, *11*, e0165521. 10.1371/journal.pone.0165521.

[13] Mohan Viswanathan, P., Sabarathinam, C., Karuppannan, S. *et al*. (2022). Determination of vulnerable regions of SARS-CoV-2 in Malaysia using meteorology and air quality data. *Environ Dev Sustain*., *24*, 8856-8882 (2022). https://doi.org/10.1007/s10668-021-01719-z.

[14] Pal, M., Berhanu, G., Desalegn, C., & Kandi, V. (2020). Severe acute respiratory syndrome Coronavirus-2 (SARS-CoV-2). https://doi.org/10.7759/cureus.7423.

[15] Prata, D. N., Rodrigues, W., Bermejo, P. H. (2020). Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. *Sci Total Environ*., *729*, 138862. Doi: 10.1016/j.scitotenv.2020.138862.

[16] Rashed, E. A. & Hirata, A. (2021). One-year lesson: Machine learning prediction of COVID-19 positive cases with meteorological data and mobility estimate in japan. *Int. J. Environ. Res. Public Health*, *18*, 5736. https://doi.org/10.3390/ ijerph18115736.

[17] Rastogi, M. (2020). Tutorial on LSTMs: A Computational Perspective. https://towardsdatascience.com/tutorial-on-lstm-a-computational-perspective-f3417442c2cd#0d00.

[18] Rendana, M & Idris, W. M. R. (2021). New COVID-19 variant (B.1.1.7). Forecasting the occasion of virus and the related meteorological factors, *Journal of Infection and Public Health*, *14*, 1320-1327. https://doi.org/10.1016/j.jiph.2021.05.019.

[19] Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On B. W., Aslam, W., *et al.* (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access*, *8*, 101489-99. https://ieeexplore.ieee.org/abstract/ document/9099302.

[20] Shrivastav, L. K., Jha, S. K. (2021). A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. *Applied Intelligence, 51*, 2727-2739. https://doi.org/10.1007/s10489-020-01997-6.

[21] Suhaimi, N. F., Jalaludin, J., Latif, M. T. (2020). Demystifying a possible relationship between COVID-19, air quality and meteorological factors: Evidence from Kuala Lumpur, Malaysia. *Special Issue on COVID-19 Aerosol Drivers, Impacts and Mitigation (III) Aerosol and Air Quality Research, 20*, 1520-1529.

[22] Talib, D., Dimitris, D., (2020). Weather impact on airborne coronavirus survival. *Physics of Fluids*, *32*(9), 093312. https://doi.org/10.1063/5.0024272.

[23] Valsamatzi-Panagiotou, A., Penchovsky, R. (2021). Environmental factors influencing the transmission of the coronavirus 2019: A review. *Environ Chem Lett., 20*, 1603-1610. https://doi.org/10.1007/s10311-022-01418-91.

[24] World Health Organization. (2020). Modes of transmission of virus causing COVID-19: Implications for IPC precaution recommendations. https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations.